COPYRIGHT © 2002 BY THE JOURNAL OF BONE AND JOINT SURGERY, INCORPORATED

Current Concepts Review

User's Guide to the Orthopaedic Literature: How to Use a Systematic Literature Review

BY MOHIT BHANDARI, MD, MSC, GORDON H. GUYATT, MD, MSC, VICTOR MONTORI, MD, P.J. DEVEREAUX, MD, AND MARC F. SWIONTKOWSKI, MD

Investigation performed at the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; Department of Medicine, Mayo Clinic, Rochester, Minnesota; and Department of Orthopaedic Surgery, University of Minnesota, Minneapolis, Minnesota

- Investigators who perform a systematic review address a focused clinical question, conduct a thorough search of the literature, apply inclusion and exclusion criteria to each potentially eligible study, critically appraise the relevant studies, conduct sensitivity analyses, and synthesize the information to draw conclusions relevant to patient care or additional study.
- A meta-analysis is a quantitative (or statistical) pooling of results across eligible studies with the aim of increasing the precision of the final estimates by increasing the sample size.
- The current increase in the number of small randomized trials in orthopaedic surgery provides a strong argument in favor of meta-analysis; however, the quality of the primary studies included ultimately reflects the quality of the pooled data from a meta-analysis.

The conduct and publication of systematic reviews of the orthopaedic literature, which often include statistical pooling or meta-analysis, are becoming more common. This article is the third in a series of guides evaluating the validity of the surgical literature and its application to clinical practice. It provides a set of criteria for optimally interpreting systematic literature reviews and applying their results to the care of surgical patients.

Authors of traditional literature reviews provide an overview of a disease or condition or one or more aspects of its etiology, diagnosis, prognosis, or management, or they summarize an area of scientific inquiry. Typically, these authors make little or no attempt to be systematic in formulating the questions that they are addressing, in searching for relevant evidence, or in summarizing the evidence that they consider. Medical students and clinicians seeking background information nevertheless often find these reviews very useful for obtaining a comprehensive overview of a clinical condition or area of inquiry.

When traditional expert reviewers make recommendations, they often disagree with one another, and their advice frequently lags behind, or is inconsistent with, the best available evidence. Reasons for disagreement among experts, and for recommendations that are inconsistent with the evidence, include a lack of attention to systematic approaches to collecting and summarizing the evidence. An evidence-based approach to surgery incorporates the patient's circumstances or predicaments, identifies knowledge gaps and frames questions to fill those gaps, includes efficient literature searches, and includes critical appraisal of the research evidence and application of that evidence to patient care. The practice of

This article is the third in a series designed to help the orthopaedic surgeon use the published literature in practice. In the first article in the series, we presented guidelines for making a decision about therapy and focused on randomized controlled trials. In the second article, we focused on evaluating nonrandomized studies that present information about a patient's prognosis. In this article, we concentrate on systematic literature reviews.

This reprint is made possible by an educational grant from Zimmer.

TABLE I User's Guide to Interpreting Review Articles

Are the results valid?
Did the review explicitly address a sensible clinical question?
Was the search for relevant studies detailed and exhaustive?
Were the primary studies of high methodological quality?
Were assessments of studies reproducible?
What are the results?
Were the results similar from study to study?
What are the overall results of the review?
How precise were the results?
How can I apply the results to patient care?
How can I best interpret the results to apply them to the care of patients in my practice?
Were all clinically important outcomes considered?
Are the benefits worth the costs and potential risks?

evidence-based medicine, therefore, is a process of lifelong self-directed learning in which caring for patients creates a need for clinically important information about diagnoses, prognoses, treatment, and other health-care issues. This article will focus on reviews that address specific clinical questions. We will provide guidelines for distinguishing a good review from a bad one and for using the results (Table I)[1,2].

Traditional reviews, or *narrative* reviews, by definition do not use a systematic approach to identifying information on a particular topic. Moreover, narrative reviews, such as those found in book chapters and instructional course lectures, often pose background-type questions and provide a general overview of a topic. An example of a background-type question is: "What are the epidemiology, clinical presentation, treatment options, and prognosis following femoral shaft fractures in adults?" We use the term *systematic review* for any summary of the medical literature that attempts to address a focused clinical question and the term *meta-analysis* for systematic reviews that use quantitative methods (i.e., statistical techniques) to summarize the results. Systematic reviews typically pose a foreground-type question. Foreground questions are more specific and provide insight into a particular aspect of management. For instance, investigators may perform a systematic review comparing the effects of plate fixation with those of nailing of humeral shaft fractures on nonunion rates (foreground question) rather than a general review of all treatments of humeral shaft fractures (background question).

When preparing a systematic review, investigators must make a host of decisions, including determining the focus; identifying, selecting, and critically appraising the relevant studies (which we will call the *primary studies);* collecting and synthesizing (either quantitatively or nonquantitatively) the relevant information; and drawing conclusions. Avoiding errors in both meta-analyses and other systematic reviews requires an organized approach, and enabling readers to assess the validity of the results of a systematic review requires explicit reporting of the methods. A number of authors have examined issues pertaining to the validity of overviews. Here, we emphasize key points from the perspective of a surgeon needing to make a decision about patient care.

Users applying the guides will find it useful to have a clear understanding of the process of conducting a systematic review (Table II). Reviewers begin by specifying the eligibility criteria for primary studies to be included in the review. Typically, reviewers identify the relevant population, intervention or exposure, and outcomes. In addition, they restrict eligibil-

TABLE II The Process of Conducting a Systematic Review

Define the question
Specify inclusion and exclusion criteria
Population
Intervention or exposure
Outcome
Methodology
Establish a priori hypotheses to explain heterogeneity
Conduct literature search
Decide on information sources: databases, experts, funding agencies, pharmaceutical companies, personal files, registries, citation lists of retrieved articles
Determine restrictions: time-frame, unpublished data, language
Identify titles and abstracts
Apply inclusion and exclusion criteria
Apply inclusion and exclusion criteria to titles and abstracts
Obtain full articles for eligible titles and abstracts
Apply inclusion and exclusion criteria to full articles
Select final eligible articles
Assess agreement between reviewers on study selection
Abstract data
Abstract data on participants, interventions, comparison interventions, study design
Abstract results data
Assess methodological quality
Assess agreement between reviewers on validity assessment
Conduct analysis
Determine method for pooling of results
Pool results (if appropriate)
Decide on handling missing data
Explore heterogeneity
Sensitivity and subgroup analysis
Explore possibility of publication bias

This reprint is made possible by an educational grant from Zimmer.

TABLE III Potential Information Resources

The Cochrane Library (www.update-software.com)
Bandolier
Best Evidence
University of York/NHS Centre for Reviews and Dissemination
MEDLINE
EMBASE
Ovid
HIRU (Health Information Research Unit) (hiru.mcmaster.ca/)
Centre for Evidence-Based Medicine at Oxford
Evidence-based medicine
ACP Journal Club

ity to studies that meet minimal methodological standards. For instance, when they are addressing a question concerning therapy, they often include only randomized clinical trials.

Having specified their eligibility criteria, reviewers then conduct a comprehensive search that typically identifies a large number of potentially relevant titles and abstracts. The reviewers then apply their inclusion and exclusion criteria to those abstracts and eventually arrive at a smaller number of primary studies. They obtain the full articles on those studies and once again apply the inclusion and exclusion criteria.

Having completed the culling process, the reviewers assess the methodological quality of the articles and abstract the data. Statistical pooling of results across studies improves the precision of the final estimates by increasing the sample size. Prior to pooling the data statistically, investigators often identify potential sources of interstudy differences, or heterogeneity. These a priori hypotheses will be examined if heterogeneity among studies is found. Finally, they summarize the data, including, if appropriate, a quantitative (statistical) synthesis or meta-analysis.

If heterogeneity among pooled studies is found in the overall meta-analysis, investigators search for potential differences among these studies by utilizing a separate sensitivity analysis. This analysis specifically includes a search for differences in the magnitude of the effect across patients, interventions, outcomes, and methodology in an attempt to explain within-study and between-study differences in results.

Conducting a meta-analysis in orthopaedics is challenging because of the paucity of clinical trails on any single topic. However, to limit bias, investigators must endeavor to adhere strictly to methodology when performing a systematic review or meta-analysis.

Clinical Scenario

You are the junior partner of a multipartner orthopaedic practice with a busy clinical service. You frequently treat major skeletal trauma, including fractures of the lower extremities. Youeyp have found that your colleagues treat certain fractures differently. For example, for the treatment of femoral and tibial shaft fractures, some use small-diameter intramedullary nails and do not ream the canal whereas others insert larger-diameter nails after intramedullary reaming.

When you ask one of your colleagues who uses the smaller-diameter nails (without reaming) for the rationale for his choice, he replies: "Nonreamed nails preserve the endosteal blood supply to the bone and that is important for fracture-healing." He adds: "Reaming the intramedullary canal increases the risk of propagating fat emboli from the canal to the lungs, leading to respiratory problems such as ARDS [adult respiratory distress syndrome] or fat embolus syndrome, particularly in multiply injured patients."

You decide to present these arguments to another colleague who uses the large-diameter nails after prior reaming. She replies: "These are just theoretical concerns. I saw a presentation about this topic at a recent meeting. I'm sure there is lots of information on this topic in the literature. Why don't you present a summary of the information on this topic at next week's rounds?"

Intrigued by this opportunity, you accept your colleague's challenge and begin to look for relevant information.

The Search

You quickly determine, from talking with fellow residents and attending surgeons, that there have been a number of randomized trials comparing intramedullary nailing techniques involving reaming with those without reaming for the treatment of femoral and tibial shaft fractures. Realizing that your one-week deadline will not be sufficient for you to summarize all of these articles, you decide to focus your literature search on identifying any recent reviews of this topic. Being relatively proficient on the Internet, you select your favorite search site, the National Library of Medicine's PubMed at www.ncbi.nlm.nih.gov/PubMed. You type in *lower extremity* and *fracture*. This identifies 4074 documents. You narrow the search by typing *overview* as a textword search, and this identifies thirteen potentially relevant papers. You review the titles of these thirteen studies and are happy to find a systematic overview and meta-analysis of intramedullary nailing with reaming compared with intramedullary nailing without reaming for the treatment of lower-extremity long-bone fractures[3]. You retrieve this article for further review. As an alternative strategy, you could have utilized the "clinical queries" section of the PubMed database and chosen a prespecified search strategy to optimize the identification of systematic reviews.

Are the Results of This Review Valid?

Did the Review Explicitly Address a Sensible Clinical Question?

Consider a systematic overview that pooled the results of all fracture therapies (both surgical and medical) for all types of fractures to generate a single estimate of the impact on fracture union rates. No clinician would find such a review useful—he or she would conclude that it is "too broad"—and no

This reprint is made possible by an educational grant from Zimmer.

reviewer has been foolish enough to conduct such an exercise. What makes a systematic review too broad? We believe that the question that clinicians ask themselves when considering this issue is: Across the range of patients and interventions that were included, and the ways that the outcomes were measured, can I expect more or less the same magnitude of effect?

The reason clinicians would reject a review of all therapies for all fracture types is that they know that some fracture therapies are extremely effective and others are harmful. Pooling across such therapies would yield an intermediate estimate of effect that is inapplicable to either the highly beneficial or the harmful interventions. The task of the clinician, then, is to decide whether the range of patients, interventions or exposures, and outcomes makes sense. Doing so requires a precise statement of what range of patients, exposures, and outcomes the reviewers have decided to consider—in other words, what are the explicit inclusion and exclusion criteria for their review? Not only do explicit eligibility criteria facilitate the user's decision regarding whether the question is sensible, but they also make it less likely that the authors will preferentially include studies that support their own prior conclusion. Bias in the choice of articles is a problem in both systematic reviews and original reports of research.

While it might seem risky, there are good reasons to choose broad eligibility criteria. First, one of the primary goals of a systematic review, and of pooling data in particular, is to adduce a more precise estimate of the treatment effect. The broader the eligibility criteria, the greater the number of studies, the greater the number of patients, and the narrower the confidence intervals. Second, broad eligibility criteria lead to more generalizable results. If the results apply to a wide variety of patients with a wide range of injury severities, the surgeon is on strong ground when applying the findings to an individual patient.

TABLE IV Quality Assessment Checklist for Randomized Trials*

	Score *(points)*		
	Yes	Partly	No
Randomization†			
Were the patients assigned randomly?	1		0
Was randomization adequately described?	2	1	0
Was treatment group concealed to investigator?	1		0
Description of outcome measurement†			
Was the description of outcome measurement adequate?	1		0
Was the outcome measurement objective?	2	1	0
Were the assessors blind to treatment?	1		0
Inclusion/exclusion criteria†			
Were inclusion/exclusion criteria well defined?	2	1	0
Were the number of excluded patients and reasons for exclusion provided?	2	1	0
Description of treatment†			
Was the therapy fully described for the treatment group?	2	1	0
Was the therapy fully described for the controls?	2	1	0
Statistics‡			
Was the test stated and a p value given?	1		0
Was the statistical analysis appropriate?	2	1	0
If the trial was negative, were confidence intervals of post hoc power calculations performed?	1		0
Was the sample size calculated before the study?	1		0
Total			
Positive trial	20		
Negative trial	21		

*Adapted from: Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol. 1992;45:255-65. Reprinted with permission from Elsevier Science. †The total maximum score was 4 points. ‡The total maximum score was 4 points if the trial was positive and 5 points if it was negative.

This reprint is made possible by an educational grant from Zimmer.

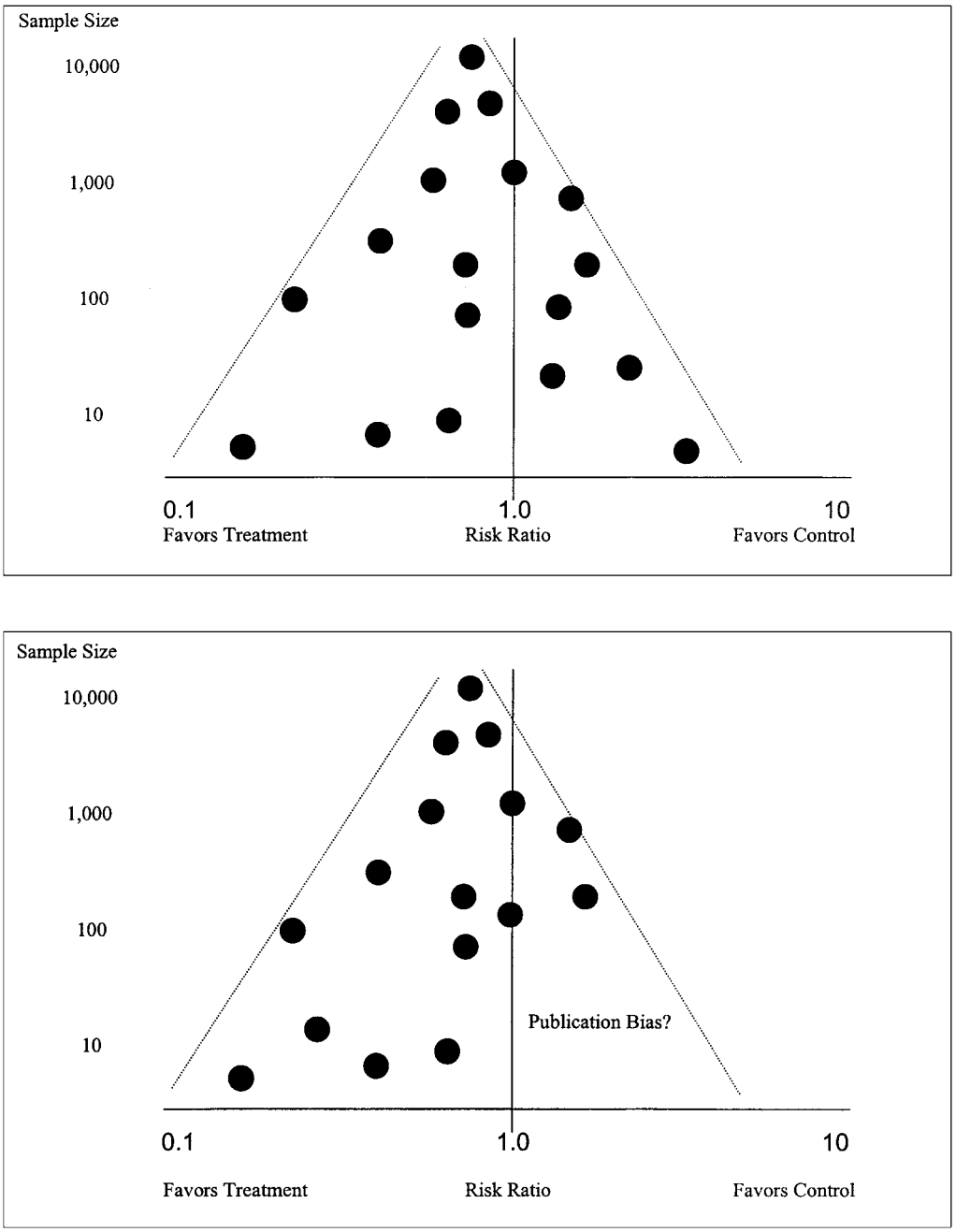

Fig. 1

Inverted funnel plot. Top panel: The sample size is plotted against the treatment effect. No evidence of publication bias exists when smaller studies with larger variability are included. Bottom panel: If small negative trials with large variances are not included, the plot will appear asymmetrical, suggesting publication bias against such negative trials.

At the same time, broad eligibility criteria leave doubt as to whether the question is sensible—i.e., they leave uncertainty as to whether the same magnitude of effect can more or less be expected across the range of patients, interventions, and outcomes.

How can reviewers resolve these conflicting demands both to generate precise and generalizable estimates of effect and, on the other hand, to avoid pooling populations or interventions that are not really comparable? One approach is to pool widely but, before beginning the review, to make a priori postulates concerning possible explanations for variability in study results. Reviewers can then test the extent to which the a priori hypotheses explain study-to-study differences in treatment effect.

Our systematic review of fracture nailing with and without reaming[3] provides a good example of this approach. The review pooled results from randomized trials addressing femo-

This reprint is made possible by an educational grant from Zimmer.

ral and tibial fractures as well as open and closed fractures. Tibial fractures differ biologically from femoral fractures in that they do not have a circumferential soft-tissue envelope that provides, in part, the blood supply to the bone, whereas an intact soft-tissue envelope around the femur is adequate to maintain blood supply to the bone and promote fracture-healing following intramedullary reaming. Thus, one might anticipate more problems when the reaming technique is used for tibial fractures. Similarly, one might anticipate that the results of reaming will be poorer for open fractures than for closed fractures, as substantial soft-tissue damage and periosteal stripping are likely to impair blood supply to the bone. These considerations raise serious questions about whether we pooled too widely when reviewing the impact of alternative nailing strategies for long-bone fractures of the lower extremities.

We were well aware of these issues. Prior to our literature search, we developed hypotheses regarding potential sources of heterogeneity. We hypothesized that heterogeneity in study results might be due to differences in the populations (the degree of soft-tissue injury [open versus closed fractures] or the type of bone [tibia versus femur]). In addition, we postulated that methodological features (quality scores and completeness of follow-up) or whether studies were published or unpublished might explain study-to-study differences in results.

Was the Search for Relevant Studies Detailed and Exhaustive?

It is important that authors conduct a thorough search for studies that meet their inclusion criteria. Their search should include the use of bibliographic databases, such as MEDLINE, EMBASE, and the Cochrane Controlled Trials Register (containing more than 250,000 randomized clinical trials); checking of the reference lists of the articles that they retrieve; and personal contact with experts in the area (Table III). It may also be important to examine books of recently published abstracts presented at scientific meetings as well as less frequently used databases, including those that summarize doctoral theses. Listing these sources, it becomes evident that a MEDLINE search alone will not be satisfactory. Previous meta-analyses in orthopaedics have variably included a comprehensive search strategy[4].

Unless the authors tell us what they did to locate relevant studies, it is difficult to know how likely it is that relevant studies were missed. There are two important reasons the authors of a review should personally contact experts in the field. The first is so that they can identify published studies that might have been missed (including studies that are in press or not yet indexed or referenced). The second is so that they can identify unpublished studies. Although some controversy about including unpublished studies remains[1,2,5,6], their omission increases the chances that studies with positive results will be overrepresented in the review (as a result of publication bias). The tendency for authors to differentially submit, and journals to differentially accept, studies with positive results constitutes a serious threat to the validity of systematic reviews.

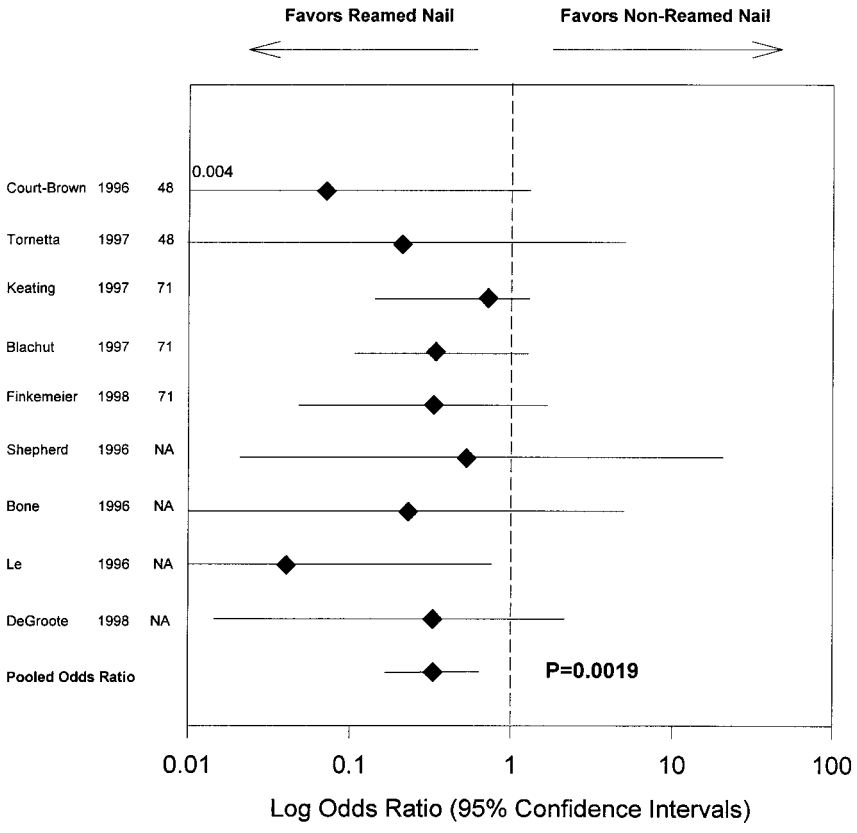

Fig. 2

Nonunion rates after treatment with intramedullary nailing with reaming. In a pooled analysis of nine randomized trials including a total of 646 patients, nailing with reaming significantly reduced the risk of nonunion compared with nailing without reaming[3]. Pooling of data is justified by widely overlapping confidence intervals, similar point estimates, and nonsignificant results of tests of heterogeneity.

This reprint is made possible by an educational grant from Zimmer.

If investigators include unpublished studies in an overview, they should obtain full written reports. They should appraise the validity of both published and unpublished studies, and they may use statistical techniques to explore the possibility of publication bias. Overviews based on a small number of small studies with weakly positive effects are the most susceptible to publication bias[2,7]. The assessment of potential publication bias can be explored visually with use of an inverted funnel plot[2]. This method uses a scatterplot of studies that relates the magnitude of the treatment effect to the weight of the study. An inverted, funnel-shaped, symmetrical appearance of dots suggests that no study has been left out, whereas an asymmetrical appearance of dots, typically in favor of positive outcomes, suggests the presence of publication bias (Fig. 1).

In our systematic review of alternative nailing strategies[3], we identified articles with MEDLINE and SciSearch and with manual hand searches of four orthopaedic journals, two textbooks, and proceedings of the annual orthopaedic meetings. We also contacted content experts. Ultimately, we identified nine randomized clinical trials (with a total of 646 patients), of which four had been published and five had not. We obtained complete manuscripts for two of the five unpublished trials. The rigor of our search methods reassure the clinician that omission of important studies is unlikely.

Were the Primary Studies of High Methodological Quality?

Even if a review article includes only randomized clinical trials, it is important to know whether they were of good quality. Unfortunately, peer review does not guarantee the validity of published research. For the same reason that our guides for using original reports of research recommend that one begins by asking if the results are valid, it is essential to consider the validity of primary articles in systematic reviews. Differences in study methods might explain important differences among the results[8]. For example, studies with less rigorous methodology tend to overestimate the effectiveness of the intervention[8,9]. Consistent results are less compelling if they come from weak studies than if they come from strong studies. Consistent results from observational studies are particularly suspect. Physicians may systematically select patients with a good prognosis to receive the therapy, and this pattern of practice may be consistent over time and geographic setting. There is no one correct way to assess validity. Some investigators use long checklists to evaluate methodological quality (Table IV), whereas others focus on three or four key aspects of the study[10-13]. Whether assessors of methodological quality should be blinded remains a subject of continued debate[13,14]. In an independent assessment of seventy-six randomized trials, Clark et al. did not find that blinding reviewers with regard to the authors or the journal in which the trials appeared significantly affected their scoring of the quality of those trials[14].

Three of the authors of our review of lower-extremity nailing independently assessed the methodological quality of each study with use of a broad-domains approach (assessment of categories of randomization and blinding, population, intervention, outcomes, follow-up, and statistical analysis) and a quality scale. The quality scores of the studies ranged from 48 to 71 points (maximum, 100 points). That approach, while rigorous, omits one important aspect of validity. Randomization may fail to achieve its purpose of producing groups with comparable prognostic features if those enrolling patients are aware of the arm to which they will be allocated. For instance, in a randomized trial comparing open and laparoscopic appendectomy, the residents responsible for enrolling patients avoided recruiting patients into the laparoscopic appendectomy group at night[2]. To the extent that patients coming in at night were sicker, this practice would have biased the results in favor of the laparoscopic appendectomy group. Concealment (i.e., ensuring that study investigators do not know the treatment to which the next patient will be allocated) is a particularly important issue in surgical trials. As it turns out, not one of the trials considered in our systematic review[3] instituted safeguards to ensure concealed randomization.

Were Assessments of Studies Reproducible?

As we have seen, authors of review articles must decide which studies to include, how valid they are, and which data to extract from them. Each of these decisions requires judgment by the reviewers, and each is subject to both mistakes (random errors) and bias (systematic errors). Having two or more people participate in each decision guards against errors, and, if there is good chance-corrected agreement between the reviewers, the clinician can have more confidence in the results of the overview[15,16].

In our systematic review comparing reaming and nonreaming techniques for nailing[3], we assessed the reproducibility of the identification and assessment of study validity with use of the kappa statistic and intraclass correlations. The kappa for the identification of potentially eligible studies was high (0.88 [95% confidence interval, 0.82 to 0.94]). The intraclass correlation coefficient for rating of study quality was also very high (0.89 [95% confidence interval, 0.73 to 0.99).

Summary of the Validity of the Meta-Analysis of Intramedullary Nailing of Long-Bone Fractures with and without Reaming

The review[3] specified explicit eligibility criteria. We are concerned that we may have pooled too broadly, given the potential differences in the relative impact of reaming compared with no reaming for nailing of femoral fractures compared with tibial fractures and of open fractures compared with closed fractures. However, we specified a priori hypotheses related to fracture site and severity. Our search strategy was comprehensive and reproducible. The studies that we found have serious methodologic limitations. However, given that they were all randomized trials, the results merit serious consideration.

What Are the Results?

Were the Results Similar from Study to Study?

We have argued that the fundamental assumption of a systematic review, and of a meta-analysis in particular, is that more or less the same magnitude of effect is anticipated across the

This reprint is made possible by an educational grant from Zimmer.

range of patients, interventions, and ways of measuring outcome. We have also noted that the goals of increasing the precision of estimates of treatment effect and the generalizability of results provide reviewers with strong, legitimate reasons for selecting relatively wide eligibility criteria. As a result, most systematic reviews document important differences in patients, exposures, outcome measures, and research methods from study to study.

Fortunately, investigators can address this unsatisfactory situation by presenting their results in a way that allows clinicians to check the validity of the initial assumption—i.e., did the results prove similar from study to study? The remaining challenge, then, is to decide how similar is similar enough.

There are three criteria to consider when deciding whether the results are sufficiently similar to warrant a single estimate of treatment effect that applies across the populations, interventions, and outcomes. First, how similar are the best estimates of the treatment effect (that is, the point estimates) from the individual studies? The more different they are, the more clinicians should question the decision to pool across studies.

Second, to what extent do the confidence intervals overlap? The greater the overlap among confidence intervals of different studies, the more powerful the rationale for pooling across those studies. One can also look at the point estimates of each individual study and determine if the confidence interval around the pooled estimate includes each of the primary point estimates.

Finally, reviewers can test the extent to which differences among the results of individual studies are greater than would be expected if all studies were measuring the same underlying effect and the observed differences were due only to chance. The statistical analyses that are used to conduct this test are called *tests of heterogeneity*[17]. When the p value associated with the test of heterogeneity is small (e.g., <0.05), chance becomes an unlikely explanation for the observed differences in the size of the effect. Unfortunately, a higher p value (0.1, or even 0.3) does not necessarily rule out important heterogeneity because, when the number of studies and their sample sizes are both small, the test of heterogeneity is not very powerful. Hence, large differences in the apparent magnitudes of the treatment effects—that is, the point estimates—among studies dictate caution in interpreting the overall findings, even in the face of a nonsignificant result of the test of homogeneity[17]. Conversely, if the differences in results across studies are not clinically important, then heterogeneity is of little concern, even if it is significant.

Reviewers should try to explain between-study differences by looking for apparent explanations (i.e., by performing sensitivity analyses). These differences include those between patients (open compared with closed fractures), between interventions (nails may be beneficial, but plates may be harmful), outcome measurements (nailing with reaming may be beneficial in promoting fracture-healing late but not early), or methodologies (the effect may be smaller in blinded trials or in those with more complete follow-up).

What Are the Overall Results of the Review?

In clinical research, investigators collect data from individual patients. In systematic reviews, investigators collect data from individual studies rather than patients. Reviewers must also summarize these data and, increasingly, they are relying on quantitative methods to do so.

Simply comparing the number of positive studies to the number of negative studies is not an adequate way to summarize the results. With this sort of approach, large and small studies are given equal weights and (unlikely as it may seem) one investigator may interpret a study as positive while another may interpret it as negative. For example, a clinically important effect that is not significant could be interpreted as positive in light of clinical importance and negative in light of significance[18]. There is a tendency to overlook small but clinically important effects if studies with nonsignificant (but potentially clinically important) results are counted as negative. Moreover, a reader cannot tell anything about the magnitude of an effect from a vote count, even when studies are appropriately classified with use of additional categories for studies with a positive or negative trend.

Typically, meta-analysts weight studies according to their size, with larger studies receiving more weight[1]. Thus, the overall results represent a weighted average of the results of the individual studies. Occasionally, studies are also given more or less weight depending on their quality, or poorer-quality studies might be given a weight of zero (i.e., they may be excluded) either in the primary analysis or in a secondary analysis that tests the extent to which different assumptions lead to different results (a sensitivity analysis). A reader should assess the overall results of an overview in the same way that he or she assesses the results of primary studies. In a systematic review of a therapy, one should look for the relative risk and relative risk reduction, or the odds. In overviews regarding diagnosis, one should look for summary estimates of the likelihood ratios.

Sometimes the outcome measures used in different studies are similar but not exactly the same. For example, different investigators might measure functional status with use of different instruments. Even if the patients and the interventions are reasonably similar, it might still be worthwhile to estimate the average effect of the intervention on functional status. One way of doing this is to summarize the results of each study as an effect size. The effect size is the difference in outcomes between the intervention and control groups divided by the standard deviation. The effect size summarizes the results of each study in terms of the number of standard deviations of difference between the intervention and control groups. Investigators can then calculate a weighted average of effect sizes from studies that measured an outcome in different ways.

Readers are likely to find it difficult to interpret the clinical importance of an effect size. (If the weighted average effect is one-half of a standard deviation, is this effect clinically trivial or large?). Once again, one should look for a presentation of the results that conveys their practical im-

This reprint is made possible by an educational grant from Zimmer.

portance (e.g., by translating the summary effect size back into conventional units). For instance, if surgeons have become familiar with the relevance of differences in functional outcome scores on a particular questionnaire, such as the Musculoskeletal Function Assessment instrument[19], investigators can convert the effect size back into differences in the scores on this particular questionnaire. Although it is generally desirable to have a quantitative summary of the results of a review, it is not always appropriate. If pooling proves inappropriate, investigators should still present tables or graphs that summarize the results of the primary studies, and their conclusions should be cautious.

How Precise Were the Results?

In the same way that it is possible to estimate the average effect across studies, it is possible to estimate a confidence interval around that estimate—i.e., a range of values with a specified probability (typically 95%) of including the true effect.

Results of the Meta-Analysis of Intramedullary Nailing of Long-Bone Fractures with and without Reaming

We tested the appropriateness of pooling data from nine trials by examining trial-to-trial variability in the results[3]. When examining our primary outcome of nonunion (Fig. 2) and implant failure rates, we found essentially similar point estimates, widely overlapping confidence intervals, and a nonsignificant result of the test of heterogeneity ($p > 0.1$). However, we also conducted a series of secondary analyses (sensitivity analyses) to explore our most questionable pooling decisions: pooling across fracture sites (femur or tibia), soft-tissue severity (open or closed fracture), publication status (published or unpublished), completeness of follow-up, and study quality score (<50 or ≥50). Although we did not find significant differences in any of these comparisons, we did find some appreciable trends. In particular, nailing with reaming was associated with a larger reduction in the rate of nonunion or implant failure in the femur (relative risk reduction, 76%) than in the tibia (relative risk reduction, 54%), nailing with reaming was associated with a larger reduction in the occurrence of the primary outcome after treatment of closed fractures (relative risk reduction, 71%) than after treatment of open fractures (relative risk reduction, 25%), and studies of lower quality showed a larger effect (relative risk reduction, 86%) than studies of higher quality (relative risk reduction, 47%). We will return to the implications of these trends toward varying effect sizes in different sorts of studies in our subsequent discussion.

In the pooled analysis across all studies, nailing with reaming was found to reduce the risk of nonunion by 67% (95% confidence interval, 32% to 84%) and to reduce the risk of implant failure by 70% (95% confidence interval, 50% to 92%) (Fig. 2). In addition, nailing with reaming did not significantly increase the risk of malunion, pulmonary complications, compartment syndrome, or infection.

How Can I Apply the Results to Patient Care?

How Can I Best Interpret the Results to Apply Them to the Care of Patients in My Practice?

The results of the systematic review of lower-extremity nailing[3] left us with some troubling apparent differences between subgroups. The reduction in the rate of adverse events associated with nailing with reaming was larger for femoral fractures than for tibial fractures, larger for closed fractures than for open fractures, and larger in poor-quality studies than in higher-quality studies. What is one to make of these trends?

Even if the true underlying effect is identical in each of a set of studies, chance will ensure that the observed results differ. As a result, reviewers risk capitalizing on the play of chance. Perhaps the studies of older patients—or, in this case, those that addressed tibial fractures—happened, simply by chance, to be those with smaller treatment effects. The reviewer may erroneously conclude that the treatment is less effective in the elderly or in those with tibial fractures. How is the reader to decide whether to believe the subgroup differences (in this case, between femoral and tibial fractures, open and closed fractures, and high and low-quality studies)?

The clinician can apply a number of criteria to distinguish subgroup analyses that are credible from those that are not. First, conclusions that are drawn on the basis of between-study comparisons (comparing patients in one study with patients in another) are less secure than those from within-study comparisons.

Other criteria that make a hypothesized difference in subgroups more credible include a big difference in treatment effect; a highly significant difference in treatment effect (the lower the p value for the comparison of the different effect sizes in the subgroups, the more credible the difference); a hypothesis that was made before the study began and was one of only a few hypotheses that were tested[20]; consistency across studies; and indirect evidence in support of the difference (biological plausibility)[1]. If these criteria are not met, the results of a subgroup analysis are less likely to be trustworthy, and one should assume that the overall effect across all patients and all treatments, rather than the subgroup effect, applies to the patient being treated and to the treatment under consideration.

All of the subgroup analyses in the nailing meta-analysis[3] were based on between-study comparisons, and none of the findings reached conventional levels of significance. These considerations suggest that differences may well have been due to chance. On the other hand, the magnitude of the differences was, in each case, substantial. In addition, we formulated our hypotheses before conducting our analysis, we tested a relatively small number of such hypotheses, and each hypothesis rested on a relatively strong biological rationale. Thus, we are left with the lingering suspicion that these subgroup differences may be real.

Were All Clinically Important Outcomes Considered?

While it is a good idea to look for focused review articles because they are more likely to provide valid results, this does not

This reprint is made possible by an educational grant from Zimmer.

mean that one should ignore outcomes that are not included in a review. For example, the potential benefits and harm of intramedullary nailing with reaming include reduced risk of nonunion and implant failure and increased risk of infection. Focused reviews of the evidence of individual outcomes are more likely to provide valid results, but a clinical decision requires consideration of all outcomes[21]. It is not unusual for systematic reviews to neglect the adverse effects of therapy.

Are the Benefits Worth the Costs and Potential Risks?

Finally, when making recommendations to their patients, surgeons must weight, either explicitly or implicitly, the expected benefits against the potential harm and cost. For example, a patient may benefit from decreased risk of infection with cast treatment of an Achilles tendon rupture at the cost (i.e., potential harm) of an increased risk of rerupture. A valid review article provides the best possible basis for quantifying the expected outcomes, but these outcomes still must be considered in the context of the patient's values and preferences about the expected outcomes of a decision[2].

Resolution of the Scenario

Our meta-analysis of intramedullary nailing of lower-extremity long-bone fractures with and without reaming[3] met most of the criteria for study validity, including explicit eligibility criteria, a comprehensive search strategy, and assessment and reproducibility of study validity[2]. However, we did not contact authors of the eligible studies for additional information. We found a very large benefit of nailing with reaming compared with nailing without reaming with regard to the rates of nonunion and implant failure, and we did not identify any adverse consequences of nailing with reaming. Furthermore, pooling of study results seems justified by the nonsignificant results of the tests of heterogeneity, the reasonable similarity of the results (point estimates), and the widely overlapping confidence intervals around those point estimates. The direction of trends toward a greater benefit for nailing with reaming and closed fractures is consistent with biological rationale. On the other hand, the quality of the studies was relatively poor, with the problems including a uniform failure to conceal randomization, and the poorer studies tended to yield larger effects.

Our interpretation is that the magnitude of the effect was sufficiently large for us to make the inference, despite the limitations in study quality, that nailing with reaming of femoral fractures provides substantially lower nonunion and implant failure rates. Given that the review failed to identify any adverse consequences of nailing with reaming, surgeons can confidently choose that procedure for femoral fractures. On the other hand, the conclusion that nailing with reaming is superior for tibial fractures, particularly open tibial fractures, is less secure. Overall, this systematic review provided information that will be very helpful for orthopaedic surgeons managing patients with lower-extremity fractures.

The current increase in the number of small randomized trials in the field of orthopaedic surgery provides a strong argument in favor of meta-analysis. However, it remains essential that those who are planning future meta-analyses adhere to accepted methodologies and provide the best available evidence to address sharply defined clinical questions[4]. While the quality of the primary studies will always be a major factor limiting the ability to draw valid conclusions, the quality of the meta-analysis is also important to ensure that the pooling of these results is as valid and free of bias as possible.

NOTE: This manuscript is based, in part, on: Guyatt GH, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: American Medical Association Press; 2001.

Mohit Bhandari, MD, MSc
Gordon H. Guyatt, MD, MSc
P.J. Devereaux, MD
Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Center, Room 2C12, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada. E-mail address for: M. Bhandari: bhandari@sympatico.ca

Victor Montori, MD
Department of Medicine, Mayo Clinic, 200 First Street S.W., Rochester, MN 55905

Marc F. Swiontkowski, MD
Department of Orthopaedic Surgery, University of Minnesota, Box 492, Delaware Street N.E., Minneapolis, MN 55455

The authors did not receive grants or outside funding in support of their research or preparation of this manuscript. They did not receive payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity. No commercial entity paid or directed, or agreed to pay or direct, any benefits to any research fund, foundation, educational institution, or other charitable or nonprofit organization with which the authors are affiliated or associated.

References

1. **Oxman A, Cook DJ, Guyatt GH.** User's guide to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. *JAMA.* 1994; 272:1367-71.
2. **Guyatt GH, Rennie D, editors.** *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: American Medical Association Press; 2002.
3. **Bhandari M, Guyatt GH, Tong D, Adili A, Shaughnessy SG.** Reamed versus nonreamed intramedullary nailing of lower extremity long bone fractures: a systematic overview and meta-analysis. *J Orthop Trauma.* 2000;14:2-9.
4. **Bhandari M, Morrow F, Kulkarni A, Tornetta P 3rd.** Meta-analyses in orthopaedic surgery. A systematic review of their methodologies. *J Bone Joint Surg Am.* 2001;83:15-24.
5. **Dickersin K.** The existence of publication bias and risk factors for its occurrence. *JAMA.* 1990;263:1385-9.
6. **Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr.** Publication bias and clinical trials. *Control Clin Trials.* 1987;8:343-53.
7. **Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA.** Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol.* 1992;45:255-65.
8. **Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP.** Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet.* 1998;352:609-13.
9. **Khan KS, Daya S, Jadad A.** The importance of quality of primary studies in

This reprint is made possible by an educational grant from Zimmer.

producing unbiased systematic reviews. *Arch Intern Med.* 1996;156:661-6.

10. **Cook DJ, Sackett DL, Spitzer WO.** Methodological guidelines for systematic reviews of randomized controlled trials in health care from the Potsdam Consultation on Meta-Analysis. *J Clin Epidemiol.* 1995;48:167-71.
11. **Cook DJ, Mulrow CD, Haynes RB.** Synthesis of best evidence for clinical decisions. In: Mulrow C, Cook D, editors. *Systematic reviews: synthesis of best evidence for health care decisions*. Philadelphia: American College of Physicians; 1998. p 5.
12. **Turner JA, Ersek M, Herron L, Deyo R.** Surgery for lumbar spinal stenosis. Attempted meta-analysis of the literature. *Spine.* 1992;17:1-8.
13. **Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ.** Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials.* 1996;17:1-12.
14. **Clark HD, Wells GA, Huet C, McAlister F, Salmi LR, Fergusson D, Laupacis A.** Assessing the quality of randomized trials: reliability of the Jadad scale. *Control Clin Trials.* 1999;20:448-52.
15. **Fleiss JL.** Measuring agreement between two judges on the presence or absence of a trait. *Biometrics.* 1975;31:651-9.
16. **Villar J, Carroli G, Belizan JM.** Predictive ability of meta-analyses of randomised controlled trials. *Lancet.* 1995;345:772-6.
17. **Cooper HM, Rosenthal R.** Statistical versus traditional procedures for summarizing research findings. *Pyschol Bull.* 1980;87:442-9.
18. **Breslow NE, Day DE.** In: *Statistical methods in cancer research*. Volume 1, The analysis of case-control studies. IARC Scientific Publications No. 32. Lyon, France: International Agency for Research on Cancer; 1980. Combination of results from a series of 2×2 tables; control of confounding; p136-46.
19. **Engelberg R, Martin DP, Agel J, Obremsky W, Coronado G, Swiontkowski MF.** Musculoskeletal Function Assessment instrument: criterion and construct validity. *J Orthop Res.* 1996;14:182-92.
20. **Assmann SF, Pocock SF, Enos LE, Kasten LE.** Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet.* 2000;355:1064-9.
21. **Colton C.** Statistical correctness. *J Orthop Trauma.* 2000;8:527-8.

This reprint is made possible by an educational grant from Zimmer.